Fiche TD11 : Apprentissage

Exercice 1 (**) *Algorithme ID3*

On considère un problème de classification binaire supervisée consistant à déterminer si un patient est malade du COVID19 en fonction de ses symptômes. Les données d'apprentissage sont de la forme (x, y) où x est un vecteur de trois attributs binaires. Le tableau ci-dessous présente ces données en décrivant un exemple par ligne ; la première colonne correspond à l'identifiant de l'exemple et la dernière à sa classe, appartenant à $\{+, -\}$:

Identifiant	Fièvre (F)	Toux (T)	Problèmes respiratoires (R)	Infecté (I)
1	non	non	non	$-$
2	oui	oui	oui	$+$
3	oui	oui	non	$-$
4	oui	non	oui	$+$
5	oui	oui	oui	$+$
6	non	oui	non	$-$
7	oui	non	oui	$+$
8	oui	non	oui	$+$
9	non	oui	oui	$+$
10	oui	oui	non	$+$
11	non	oui	non	$-$
12	non	oui	oui	$-$
13	non	oui	oui	$-$
14	oui	oui	non	$-$

1. Appliquer en détaillant le processus l'algorithme ID3 au jeu d'apprentissage ci-dessus.

2. Selon le modèle obtenu à la question précédente, un malade se présentant chez le médecin avec de la toux et des problèmes respiratoires mais pas de fièvre est-il malade du COVID19 ?

Exercice 2 (*) *Méthode des k plus proches voisins*

On considère un ensemble d'étudiants ayant tous la même moyenne générale mais regroupés dans deux groupes selon leur compétences. On a déjà classé 6 étudiants et on souhaite en classer un 7ème.

	Maths	Info	Français	Anglais	
Tom	20	20	0	0	G1
Eliott	8	8	12	12	G2
Amélie	20	20	0	0	G1
Jean	0	0	20	20	G2
Lily	10	10	10	10	G2
Bob	2	2	18	18	G1
Alice	9	15	13	11	?

Pour déterminer le groupe d'Alice, on utilise la méthode des k plus proches voisins avec comme distance entre deux 4-uplets de notes la distance dérivée de la norme 1. Donner une valeur de k raisonnable pour appliquer cet algorithme et déterminer un groupe pour Alice.

Exercice 3 (***) *Fléau de la dimension*

On considère un problème de classification dont les entrées sont des vecteurs de $[0, 1]^d$ pour $d \geq 1$. Cet hypercube de dimension d est partitionné en petits hypercubes de côté ε et on fait l'hypothèse que pour qu'un point inconnu x soit correctement classé, il est nécessaire et suffisant qu'il y ait au moins un point du jeu d'entraînement dans le petit hypercube de côté ε contenant x. On note n la taille du jeu d'entraînement.

1. Exprimer la probabilité qu'un élément soit correctement classé en fonction de n, d et ε.

2. En déduire la façon dont doit évoluer n par rapport à la dimension d pour espérer avoir un algorithme de classification pertinent.

Remarque : Cet exemple illustre sur une situation très simplifiée le "fléau de la dimension" (curse of dimensionality) : en grande dimension, un nuage de points est a priori très éparpillé à moins que le nombre de points dans ce nuage soit exponentiel en la dimension. Or, beaucoup d'algorithmes d'apprentissage reposent sur l'utilisation du

voisinage connu d'un point inconnu pour en déterminer la classe. Par conséquent, en grande dimension, le nombre de données nécessaires à une classification fiable doit être extrêmement grand. Pour cette raison, lorsque les entrées d'un problème de classification vivent dans un espace de grande dimension, on commence généralement par en réduire le nombre, soit en déterminant les dimensions les plus significatives via des méthodes d'apprentissage non supervisé, soit en retrouvant la variété (de dimension plus réduite) dans laquelle évoluent les points d'entraînement.

Exercice 4 (*) *Classification bien séparée*

On considère le jeu de données de points unidimensionnels suivant : $6, 14, 22, 38, 46, 54, 60$.

1. Appliquer l'algorithme des k-moyennes sur ce jeu avec $k = 2$ et en prenant comme centres initiaux 6 et 25.
2. On dit qu'une classification est bien séparée si pour tout point x appartenant à la classe C il n'existe pas de couple de points $(y, z) \in C \times \overline{C}$ tels que $d(x, y) > d(x, z)$. La classification obtenue à la question précédente est-elle bien séparée ?

Exercice 5 (***) *Terminaison de l'algorithme des k-moyennes*

Dans cet exercice, on se propose de montrer la terminaison de l'algorithme des k-moyennes. Les notations sont celles du cours. On excluera les cas pathologiques (centroïdes égaux, égalités de la distance d'un point à plusieurs centroïdes différents) ; ces derniers pouvant être évités en rajoutant quelques lourdeurs à l'algorithme.

1. Soit $x_1, ..., x_p$ des points d'un espace euclidien $(E, \langle ., . \rangle)$. Montrer que la fonction

$$f : x \longmapsto \sum_{i=1}^{p} ||x - x_i||^2$$

admet un unique minimum global sur E en le barycentre des x_i.

2. Montrer que lors de la partition de X en k classes $X_1, ..., X_k$, la quantité $\sum_{i=1}^{k} \sum_{x \in X_i} ||x - c_i||^2$ où les c_i sont les barycentres des X_i ne peut prendre qu'un nombre fini de valeurs.
3. Montrer que cette quantité décroît strictement au fil des itérations de l'algorithme des k-moyennes appliqué à X et conclure.

Exercice 6 (*) *CHA(T)*

On souhaite établir une classification des félins de sorte à mieux comprendre la genèse évolutive de cette catégorie d'animaux. Pour ce faire, l'ADN de différents individus est examiné : i_1 est un chat domestique, i_2 un puma, i_3 un lynx, i_4 un ocelot, i_5 un caracal et i_6 un jaguar. Les distances entre chacun de ces individus (caractérisant leur ressemblance quant à l'ADN), sont données dans le tableau suivant :

	i_1	i_2	i_3	i_4	i_5	i_6
i_1	0	0.84	0.95	1.49	1.85	2.56
i_2		0	0.7	1.11	1.87	2.38
i_3			0	1.35	2.05	2.15
i_4				0	1.96	2.32
i_5					0	2.30
i_6						0

Etablir une classification hiérarchique ascendante sur ce jeu de données en construisant le dendrogramme associé lorsque la distance est évaluée :

1. Selon le lien minimum.
2. Selon le lien maximum.
3. Selon le lien moyen.